Protein DNA Sequence Classification Challenge

Kolawole Tajudeen[1], Moshood O. Yekini[2]
Department of Mathematics and Computer Science
African Masters in Machine Intelligence, Accra, Ghana
ktajudeen@aimsammi.org[1], myekini@aimsammi.org[2]

1 Introduction

Modern techniques are being applied to the interpretation of genetic sequences for effectiveness and vast applications for either homogeneous or heterogeneous species identification. This process is called DNA Sequencing Analysis. With the bands of black and white images being one of the most recognizable examples which contain different fragment of DNA sorted as genetic fingerprint for different sample comparison. [1]

2 Choice of Feature Engineering

We are provided with a 2-class labelled "Bound" dataset with DNA sequences of 2000 and 1000 training and testing dataset. We extracted features from the dataset using a 3-choice approach of the listed steps:

a. taking each character of the DNA sequence as a single entity feature

b. by treating the DNA sequence as a language i.e k-mer counting

c. a preprocessing CountVectorizer with 'char_wb' analyzer of 6-ngram range

3 Experimentation

On experimentation, we built from scratch the implementation of the multinomial naive bayes, kernel logistic regression, kernel ridge regression and kernel support vector machine model algorithms. The kernel ridge regression is modified in order to cater for the classification task in its implementation and other likely parameters are tunned to achieve satifisfiable metrics of the model performance on training, validation and testing.

Table 1: Result summary of experimentation

Model	Feat. Eng.	Training Acc.	Validation Acc.	Public Score
MultinomialNaiveBayes	a	0.6890	_	0.68200
KernelSVM	b and c	0.68625	0.6750	0.6663
KernelRidgeRegression	b and c	0.7175	0.6000	0.65400
KernelLogisticRegression	b and c	0.68813	0.6075	0.6100

For the multinomial naive bayes, we decided to trained on all the dataset without splitting and later predicted which we achieved the highest score on the pubic score.
Considering the nature of the data and our choices of preprocessing, there is high likelihood of overfitting in the private testing dataset which makes us to carefully examined our approach of possible tunning of hyperparameters and the number of feature selection in the case of the CountVectorizer and ngram range in our feature engineering approach.

4 Conclusion and Recommendation

We recommend finding better techniques in feature engineering of the dataset to would yield maximum accuracy and validation score without overfitting on the test dataset. Moreso, the use well featured encoding techniques to prioritize importance features can be examined on both the matrice or sequence dataset. Thus, this will be suitable as a generalized model for deploration and due concern will be given to the metrics beyond accuracy for effective decision making.

Acknowledgments

The authors heartedly appreciate and acknowledge African Masters in Machine Intelligence (AMMI) and the lecturing team from INRIA, GoogleAI and MINES ParisTech in person of Jean-Philippe Vert, Julien Mairal, and Romain Menegaux

References

[1] https://dnatestingchoice.com/en-us/news/what-is-dna-analysis [2] https://github.com/rmenegaux/kernels-AMMI-2020

2nd Protein DNA Sequence Classification Challenge African Masters in Machine Intelligence (AMMI 2020), Accra, Ghana.